Towards Certified Ethical Artificial Intelligences

Steve Tueno

Université de Sherbrooke, Québec, Canada

Abstract

This paper is about a proposal to allow the definition of ethical artificial intelligences for which ethics can be certified. The proposal is based on definition of ethics ontologies, each ontology aiming at formalizing some ethical principles of interest. During the learning phases of machine learning algorithms, the ethics ontologies will be exploited in order to reason on learning data and build ethically acceptable machine learning models. For instance, an ontology may describe that learning can only be achieved if the learning data is almost equitably balanced between men and women. Thus, a learning algorithm may link possible values of a coefficient to be determined to either the percentage of data for which the sex field is set to 'F' or to the ones for which the field is set to 'M'. Machine learning models generated following the proposal / algorithms / tools will be certified ethically correct, which will have an impact on their adoption and usage.

Keywords: Artificial Intelligence, Machine Learning, Ethical Correctness, Ontologies

Introduction

1. Background

References

Email address: steve.jeffrey.tueno.fotso@usherbrooke.ca (Steve Tueno)